

Lalu Kumar

AI • Full Stack Engineer

connect.lalukumar@gmail.com • +91 8789485074 • github.com/CodeForgeNet • linkedin.com/in/lalu-kumar • Portfolio

Professional Summary

AI Systems Engineer specializing in **AI-native architectures** and production-grade full-stack systems. Expert in building scalable RAG pipelines, optimizing LLM inference via **semantic caching**, and bridging high-concurrency Java/Spring Boot backends with frontier AI models. Proven track record in reducing operational costs through automated evaluation frameworks and efficient data ingestion strategies.

Technical Skills

AI Engineering: LangChain, RAG Pipelines, Semantic Search (Vector DBs: Pinecone), Semantic Caching (Upstash Redis), Anthropic MCP, OpenAI/Gemini API Optimization

Full Stack Development: Java (Spring Boot), Node.js (NestJS), TypeScript, Next.js, React, JPA/Hibernate, Lombok, MapStruct

Infrastructure: AWS (Lambda, S3, EC2), Docker, Serverless Framework, CI/CD (GitHub Actions), SQLite, MySQL

Low-Level & Real-Time: TensorFlow.js, Socket.io, IMAP Protocol, NPM Package Development

Experience

AI Systems Engineering (Independent)

Jun 2025 - Present

AI Engineer

- Engineered and published **TunePrompt**, an NPM framework for automated semantic evaluation of LLM outputs, utilizing SQLite for versioned prompt history.
- Implemented **Upstash Redis Semantic Caching** for RAG systems, resulting in a **25% reduction in API token expenditure** and achieving sub-50ms response times for cached queries.
- Developed a streamlined **data ingestion pipeline** for RAG applications, implementing specialized chunking strategies and metadata filtering to improve context retrieval precision.

QSpiders Development Center

Oct 2024 - May 2025

Full Stack Engineer (Associate)

- Developed high-concurrency RESTful services using **Spring Boot** and **JPA**, optimizing database query performance for MySQL and Oracle environments.
- Reduced backend development overhead by **40%** through the implementation of generic repository patterns and automated DTO mapping using **Lombok**.
- Architected real-time state management systems in React, ensuring seamless UI updates for data-intensive enterprise applications.

Selected Projects

TunePrompt - LLM Observability & Evaluation Framework

[GitHub]

Node.js, TypeScript, OpenAI/Anthropic APIs, SQLite

- Architected a local-first CLI tool for prompt engineering that utilizes **Cosine Similarity** to benchmark LLM outputs against reference datasets, enabling objective performance tracking.
- Engineered a **CI-native testing engine** to automate prompt regression checks, preventing the deployment of models that fail to meet semantic accuracy thresholds.
- Designed an **SQLite-backed history engine** to store and visualize model performance across different versions, facilitating data-driven prompt optimization.

AI-Agent System & Serverless RAG Infrastructure

[Live]

Next.js, AWS Lambda, Pinecone, Upstash Redis, React Three Fiber

- Developed a high-performance **serverless RAG backend** on AWS Lambda, achieving **95% retrieval precision** for context-aware career and technical consultations.
- Optimized system overhead by implementing a **Multi-Layer Caching strategy** with Redis, reducing LLM cold-start latency and minimizing API costs for recurring queries.
- Integrated a 3D interactive agent using **React Three Fiber**, leveraging optimistic UI updates to mask network latency and improve user engagement.

LucidGrowth - Enterprise Email Trace & Analysis System

[Live]

NestJS, Next.js 15, TypeScript, MongoDB, IMAP Protocol, Mailparser

- Engineered a **high-concurrency NestJS backend** to automate the ingestion and parsing of raw IMAP metadata, utilizing **Service-Oriented Architecture (SOA)** to decouple protocol-level fetching from data persistence.
- Developed a **modular filtering and aggregation engine** in MongoDB, enabling sub-100ms subject-based retrieval and re-verification of historical email transmission logs.
- Architected a responsive **Next.js 15 dashboard** with real-time state synchronization, visualizing technical headers and delivery statuses to streamline the debugging of complex enterprise mail flows.

Certifications

Model Context Protocol (MCP) Specialist - Anthropic (2026)

AWS Solutions Architecture - Job Simulation (2025)

Education

Master of Computer Applications (MCA) - IGNOU

Expected Jan 2026

Bachelor of Computer Applications (BCA) - MMHAPU, Patna

2023